A NOTE ON ESTIMATING PROPORTIONS BY LINEAR REGRESSIONS

Alvin A. Cook, Jr.

October 1971

-4712

8R

A NOTE ON ESTIMATING PROPORTIONS BY LINEAR REGRESSION

Alvin A. Cook, Jr.*

The RAND Corporation, Santa Monica, California

The estimation of the parameters of a linear regression model in which the dependent variable is a fraction or proportion frequently occurs in statistical analyses. This proportion reflects some specific activity such as the proportion of income spent on some good and is related to a number of exogenous characteristics. The estimation of a proportion relationship is not always performed in isolation. The relationship of the exogenous variables to the remainder of the population, or complementary proportions, is usually of equal interest.

An example of such an analysis has been performed by McCall[1] in studying the movements of people into and/or out of various income classes over time. In one part of his study McCall calculates a first order Markov transition matrix of peoples' movements into, out of, or remaining in income classes. Each cell contains a transition probability reflecting the propensity of individuals to move out of a given income class or remain in it. The changes in the transition probabilities for each cell are then related to changes in GNP for the period

---

*Any views expressed in this paper are those of the author. They should not be interpreted as reflecting the views of The RAND Corporation or the official opinion or policy of any of its governmental or private research sponsors. Papers are reproduced by The RAND Corporation as a courtesy to members of its staff.

[1]John J. McCall, Earnings Mobility and Economic Growth, R-526-OEO, The RAND Corporation, October 1970.

from 1958 to 1966. In each matrix the probabilities sum to one and each cell will undoubtedly be affected differently by changes in GNP.

In McCall's analysis and in many analyses, the probabilities or proportions across all cells sum to one in the raw data. For prediction purposes it is desirable that the estimated proportions always sum to one as well. In this note we show that the estimated proportions sum to one and no constraint is needed on the proportions if the parameter estimates are BLUE. If the parameters are not, then a constraint can be constructed using Zellner's[2] technique of Seemingly Unrelated Least Squares (SULS) to ensure that the estimated proportions sum to one. It is further noted that the results can be generalized to any system of equations containing the same exogenous variables in each equation and specifying an exact linear constraint on the dependent variables for all observations in the raw data.

UNCONSTRAINED LEAST SQUARES

Assume that there are z proportions p_i and n observations on each p_i such that $\sum_{i=1}^{z} p_{ij} = 1$, for every $j = 1, \ldots, n$. Each regression equation can then be written

$$p_i = X\beta_i + \varepsilon_i \qquad i = 1, \ldots, z \tag{1}$$

where each p_i' is a 1 x n vector of observations $(p_{i1}, \ldots, p_{in})$, $X = (1, X_1, \ldots, X_s)$ where each X_k is an n x 1 vector of observations, β_i is an (s+1) x 1 vector of regression coefficients, $\varepsilon_i \sim N(0,\sigma^2)$,

[2] Arnold Zellner, "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," Journal of the American Statistical Association, Vol. 57, June 1962, pp. 348-368.

$E(\varepsilon_i X_k) = 0$ for all i and k, $E(\varepsilon_{ij}\varepsilon_{ij'}) = 0$, $\sum_{i=1}^{z} p_{ij} = 1$ for all j, and $0 \leq p_{ij} \leq 1$, all i, j.

The equations in (1) represent actually z-1 independent equations and $\sum_{i=1}^{z-1} p_i = 1-p_z$. Thus p_z can be easily calculated, but the separate effects of the exogenous variables on p_z are omitted. To ensure that $\sum_{i=1}^{z} \hat{p}_i = 1$ and to obtain the separate coefficients $\hat{\beta}_{ik}$ for all i and k, it is not necessary to use a Lagrangean constraint. Estimation of (1) by Ordinary Least Squares (OLS) yields the unbiased estimate

$$\hat{\beta}_i = (X'X)^{-1}X'p_i , \qquad i = 1, \ldots, z \tag{2}$$

and

$$\hat{p}_i = X\hat{\beta}_i , \qquad i = 1, \ldots, z . \tag{3}$$

By summing the estimated proportions over all i, we obtain

$$\sum_{i=1}^{z} \hat{p}_i = \sum_{i=1}^{z} X\hat{\beta}_i = \sum_{i=1}^{z} X(X'X)^{-1}X'p_i \tag{4}$$

$$= \sum_{i=1}^{z} p_i = 1 .$$

Therefore OLS ensures that the $\hat{p}_i$ sum to one regardless of variances of each estimate as long as the $\hat{\beta}_i$ are unbiased. Similarly the variance of the sum of the estimated proportions is zero using the above result:

$$\text{Var}\left(\sum_{i=1}^{z} \hat{p}_i\right) = E\left(\sum_{i=1}^{z} \hat{p}_i - \sum_{i=1}^{z} p_i\right)^2 \tag{5}$$

$$= E\left[\left(\sum_{i=1}^{z} \hat{p}_i\right)^2 + \left(\sum_{i=1}^{z} p_i\right)^2 - 2\left(\sum_{i=1}^{z} \hat{p}_i\right)\left(\sum_{i=1}^{z} p_i\right)\right]$$

$$= 0 .$$

IMPLICIT CONSTRAINTS ON THE REGRESSION COEFFICIENTS

The fact that the proportions sum to one implies a constraint on β_{ik} across all i equations. If the β_i are BLUE (and the X_k are fixed or independently distributed), then $\sum_{i=1}^{z} p_i = 1$ implies $\sum_{i=1}^{z} \hat{\beta}_{ij} = 0$ for all j. Consider

$$(6) \quad d\left(\sum_{i=1}^{z} \hat{p}_i\right) = \sum_{i=1}^{z} d\hat{p}_i = \sum_{i=1}^{z} \sum_{j=1}^{s} \frac{\partial \hat{p}_i}{\partial X_j} dX_j$$

$$= \sum_{i=1}^{z} \sum_{j=1}^{s} \hat{\beta}_{ij} dX_j$$

$$= \sum_{j=1}^{s} \left(\sum_{i=1}^{z} \hat{\beta}_{ij}\right) dX_j$$

$$= (\hat{\beta}_{11} + \ldots + \hat{\beta}_{z1})dX_1 + \ldots + (\hat{\beta}_{1s} + \ldots + \hat{\beta}_{zs})\, dX_s$$

$$= 0 .$$

Considering a total change in the summation of the $\hat{p}_i$ due to a total change in any exogenous variable, say X_j, yields

$$(7) \quad 0 = (\beta_{11} + \ldots + \beta_{z1}) \frac{dX_1}{dX_j} + \ldots + (\beta_{1j} + \ldots + \beta_{zj}) + \ldots$$

$$+ (\beta_{1s} + \ldots + \beta_{zs}) \frac{dX_s}{dX_j} .$$

But $dX_i/dX_j = 0$ for $i = 1, \ldots, s$; $i \neq j$. By considering total changes in $\sum_{i=1}^{z} \hat{p}_i$ due to dX_j, $j = 1, \ldots, s$, we obtain the result that $\sum_{i=1}^{z} \hat{\beta}_{ij} = 0$ for every j. The summation of intercept terms, however, does not equal zero:

$$(8) \quad \hat{\beta}_{i0} = \bar{p}_i - \sum_{j=1}^{s} \hat{\beta}_{ij} \bar{X}_j$$

and

$$(9) \qquad \sum_{i=1}^{z} \hat{\beta}_{i0} = \sum_{i=1}^{z} \bar{P}_i - \sum_{i=1}^{z} \sum_{j=1}^{s} \hat{\beta}_{ij} \bar{X}_j$$

$$= 1 - \sum_{j=1}^{s} \left(\sum_{i=1}^{z} \hat{\beta}_{ij} \right) \bar{X}_j \ .$$

Based on the above result of $\sum_{i=1}^{z} \hat{\beta}_{ij} = 0$ for every j,[3]

$$(10) \qquad \sum_{i=1}^{z} \hat{\beta}_{i0} = 1 \ .$$

It is immediate from (4) that an exact linear constraint on the dependent variables in the raw data implies the same exact linear constraint for the estimated dependent variables across the z equations. Further, the results of Eqs. (6) to (10) imply that the $\hat{\beta}_{ij}$ sum to zero across i and the $\hat{\beta}_{i0}$ sum to the value of the constraint if an exact linear constraint exists on the dependent variables. Moreover, as will be seen below, the parameters may be constrained to ensure the exact linear constraint on the dependent variables if the parameters are not unbiased.

ESTIMATION WITH THE COEFFICIENT CONSTRAINT

If the parameters are not BLUE, the $\hat{P}_i$ and $\hat{\beta}_{i0}$ may not sum to one, nor may the $\sum_{i=1}^{z} \hat{\beta}_{ij}$ sum to zero for every j. These conditions can be

[3] Another proof of these results was pointed out to me by B. Efron. Let e_j represent the (jx1) vector of ones. Write all z equations $\hat{P}_i = X\hat{\beta}_i$ side by side to obtain $\hat{P} = X\hat{\beta}$ where $\hat{P}$ is n x z, X is n x (s+1), and $\hat{\beta}$ is (s+1) x z. It is then $Pe_z = X\beta e_z$. But $Pe_z = e_n$ by equation (4) and $X \cdot (1,0,\ldots,0)' = e_n$ by definition. This implies that $\beta e_z = (1,0,\ldots,0)$ which contains the above results.

forced on the system however by treating the z equations as a system of seeming unrelated regressions as developed by Zellner. Rewrite (1) as

$$(11) \qquad \begin{bmatrix} p_1 \\ \vdots \\ p_z \end{bmatrix} = \begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & & \vdots \\ & & \ddots & \\ 0 & \cdots & \cdots & X_z \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_z \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_z \end{bmatrix}$$

where each X_i is a vector of independent variables $(1, X_{i1}, \ldots, X_{is})$ and each β_i a vector $(\beta_{i0}, \beta_{i1}, \beta_{i2}, \ldots, \beta_{is})$; Eq. (10) can be simplified as

$$(12) \qquad p = X\beta + u \ .$$

Application of least squares yields the BLU estimator,

$$(13) \qquad \beta^* = (X'\Sigma^{-1}X)X'\Sigma^{-1}p \ ,$$

where Σ is estimated by the disturbance variance-covariance matrix.[4] By further constraining the β_{ij} to sum to zero across all i equations for every j and the β_{i0} to sum to one, we obtain simultaneous constraint estimates[5]

$$(14) \qquad \tilde{\beta} = \beta^* + (X'\Sigma^{-1}X)^{-1}Q'[Q(X'\Sigma^{-1}X)^{-1}Q']^{-1}(W-O\beta^*) \ .$$

---

[4] See A. Zellner, "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," Journal of the American Statistical Association, June 1962.

[5] See H. Theil, Economic Forecasts and Policy (2nd ed.), North-Holland Publishing Co., Amsterdam, 1961.

The estimate $\tilde{\beta}$ satisfies the constraint $Q\tilde{\beta} = W$, where Q is a matrix specifying the combination of the $\tilde{\beta}_{ij}$ and W is a vector describing the constraint. For the constraints above the intercepts sum to one and the slope coefficients across each X_j sum to zero, therefore W = (1, 0, ..., 0), a 1 x (s+1) vector. Thus all constraints may be forced in the event that unconstrained OLS is not consistent with the implied constraints on the β_{ij}.